УДК 519.24

АНАЛИЗ ПРЕДСКАЗАТЕЛЬНЫХ СВОЙСТВ МОДЕЛИ В ВИДЕ РЕГРЕССИОННОГО УРАВНЕНИЯ

В.А. Качалкин,
Казанский (Приволжский) федеральный университет,
Россия, г. Казань

Ключевые слова: *регрессионная модель, регрессионный анализ, функция отклика.*

Оценка регрессионного уравнения может быть использована для предсказания некоторого промежуточного значения отклика Y. Интервальная оценка математического ожидания величины Y определяется следующим образом:

$$\hat{Y} - t_{1-\alpha/2} S_R \le Y \le \hat{Y} + t_{1-\alpha/2} S_R \ . \qquad (1)$$

Для уравнения $$Y = b_0 + b_1(x - \bar{x}) \qquad (2)$$

схематически изображена линия регрессии и её доверительный интервал (Рис.1).

Геометрическое место точек, соответствующих доверительным пределам, представляет собой две кривые, расстояние между которыми минимально в точке $\bar{x}$. Результаты наблюдений отклика в каждой точке претерпевают рассеивание вблизи поверхности отклика вследствие действия случайных неконтролируемых факторов. Величина этого рассеивания характеризуется дисперсией воспроизводимости эксперимента $\sigma^2\{y\}$. С дру-

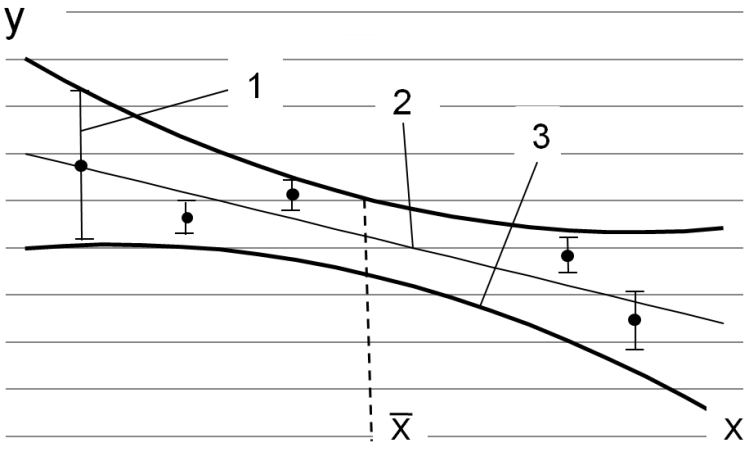

Рис.1. Оценка линии регрессии и доверительный интервал для Y
1- доверительный интервал для отдельного измерения Y при x_i; 2- оценка линии регрессии; 3- геометрическое место доверительных пределов для Y

гой стороны, те же результаты рассеяны относительно найденной оценки $\hat{Y}(\vec{X},\vec{B})$ регрессионной модели, приближающей неизвестную функцию отклика, причём это рассеивание определяется не только случайной погрешностью опыта, но и возможной систематической ошибкой в случае неправильно сделанного предположения о форме регрессионной модели. Такое рассеивание характеризуется остаточной дисперсией S_R^2 . Если вид регрессионной модели выбран адекватно функции отклика, то остаточная дисперсия является несмещённой оценкой дисперсии воспроизводимости.

Если же вид регрессионной модели выбран неадекватно функции отклика, то рассеивание наблюдений относительно оценки регрессионной модели больше рассеивания их относительно функции отклика за счёт влияния помимо случайных величин неконтролируемых факторов ещё и указанной неадекватности. Это рассеивание характеризуется остаточной дисперсией, статистически значимо отличающейся от дисперсии воспроизводимости.

Так как значение $\sigma^2\{y\}$, как правило, неизвестно, следует использовать её оценку $S^2\{y\}$, найденную по наблюдениям параллельных опытов.

Проверка предпосылки об адекватности регрессионной модели и функции отклика заключается в сравнении остаточной S_R^2 и выборочной $S^2\{y\}$ дисперсий воспроизводимости. Однородность этих дисперсий свидетельствует о том, что остаточная дисперсия может быть оценкой генеральной дисперсии воспроизводимости и о том, что предположение об адекватности выбранной формы регрессионной модели и неизвестной функции отклика не противоречит данным регрессионного эксперимента.

В качестве критерия проверки предположения об однородности указанных дисперсий используется дисперсионное отношение

$$F = \frac{S_R^2}{S^2\{y\}} \ . \tag{3}$$

Если вычисленное значение F-критерия получается меньше $F_{1-\alpha}(\nu_R,\nu)$, определяемого по таблицам или по встроенным статистическим функциям табличного процессора MS Excel для выбранного уровня значимости α и степеней свободы ν_R, ν, то гипотезы об однородности сравниваемых дисперсий, а следовательно, и об адекватности регрессионной модели и функции отклика не отвергаются, и наоборот.

При проведении пассивного эксперимента, если $S^2\{y\}$ неизвестна, то для проверки пригодности полученного уравнения регрессии для прогнозирования значений отклика используется дисперсионное отношение

$$F_{расч} = \frac{S_y^2}{S_R^2}, \tag{4}$$

где S_y^2 - дисперсия рассеяния отклика относительно среднего значения,

$$S_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2, \qquad (5)$$

здесь y_i - текущее значение отклика; $\bar{y}$ - среднее значение отклика.

Значение F-критерия, рассчитанное по формуле (4), сравнивается с табличным значением $F_{табл}$ для выбранного уровня значимости α и степеней свободы $\nu_1 = N-1,\ \nu_2 = N-k-1$.

Для того, чтобы уравнение модели можно было считать удовлетворительным для целей прогноза (в том смысле, что размах предсказываемых значений отклика значительно больше, чем стандартная ошибка отклика), значение $F_{расч}$, рассчитанное по формуле (4), должно не просто превышать выбранную процентную точку F-распределения, а превосходить её в несколько раз (примерно в 4 раза).

При выборе формы регрессионной зависимости, для установления целесообразности перехода к более сложным формам, остаточные дисперсии линейной и нелинейной зависимостей сравниваются по F-критерию. Если приближение по более сложной зависимости приводит к уменьшению S_R^2, но это уменьшение несущественно по F-критерию, принимается более простая форма связи. Значение F-критерия рассчитывается по формуле

$$F_{расч} = \frac{S^2_{R_{линейн}}}{S^2_{R_{нелинейн}}}. \qquad (6)$$

Значение $F_{расч}$, рассчитанное по формуле (6), сравнивается с табличным значением для заданного уровня значимости α и степеней свободы $\nu_1 = N-k_1-1,\ \nu_2 = N-k_2-1$ (где N - количество опытов; k_1 - количество параметров линейной зависимости; k_2 - количество параметров нелинейной зависимости). Если $F_{расч} < F_{табл}$, то уменьшение S_R^2 за счёт введения нелинейности несущественно.

О полноте представления факторов объекта описываемым регрессионным уравнением можно судить по величине множественного коэффициента корреляции

$$R = \sum_{j=1}^{k} r_{yx_j} b_j^*, \qquad (7)$$

где r_{yx_j} - выборочный коэффициент корреляции между откликом и фактором x_j; b_j^* - коэффициенты уравнения регрессии в стандартизованном виде.

Значимость R может быть проверена по F-критерию

$$F = \frac{R^2}{1-R^2}\frac{N-k-1}{k} \ . \qquad (8)$$

Если $F_{расч} > F_{табл}$, то R считается существенным, в противном случае - несущественным.

Табличное значение F-критерия выбирается для заданного уровня значимости α и степеней свободы $\nu_1 = k$, $\nu_2 = N-k-1$.

Литература

1. Дубров А.М., Мхитарян В.С., Трошин Л.И. Математическая статистика для бизнесменов и менеджеров. М., МЭСИ, 2000, 140 с.

2. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. Исследование зависимостей. М., Финансы и статистика, 1985, 487 с.

3. Адлер Ю.П., Маркова Е.В., Грановский Ю.В. Планирование эксперимента при поиске оптимальных условий. М.: Наука, 1970. – 280 с.

4. Дрейпер Н., Смит Г. Прикладной регрессионный анализ. М., «Статистика», 1973.